INFORMED SOURCE SEPARATION FROM MONAURAL MUSIC WITH LIMITED BINARY TIME-FREQUENCY ANNOTATION

Il-Young Jeong[1] *and Kyogu Lee*[1, 2]

[1]Music and Audio Research Group, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Korea
[2]Advanced Institutes of Convergence Technology, Suwon, Korea

ABSTRACT

This paper presents a novel informed audio source separation algorithm given a limited binary time-frequency annotation. Assuming that all the sources can be represented using a low-rank model, we derive an objective function to minimize the rank of the source spectrogram, and the error between the target and the estimated coefficients. Especially, we apply the nuclear norm and l_1-norm, which allow a relaxation of the model, and represent them in the convex formulation. Experimental results show that the proposed method achieves better and more robust separation performance than the state-of-the-art under the incomplete and inexact annotation condition.

Index Terms— Informed source separation, limited T-F annotation, nuclear norm, augmented Lagrangian multiplier

1. INTRODUCTION

Informed source separation (ISS) is an approach to increase the separation performance using additional information about the sources. Here, this information can be provided in various forms. For example, users can guide the separation procedure by means of humming [1, 2] or tapping [3], or by providing the score [4], the contour of the fundamental frequencies [5], or the lyrics (when the target source is vocal) [6].

A time-frequency (T-F) annotation for each source is also useful information, and has proven to achieve a significant separation performance gain. Bryan *et al*. presented an interactive sound source separation method [7] using probabilistic latent component analysis [8]. Lefèvre *et al*. took a similar approach using non-negative matrix factorization [9, 10], and more recently suggested a convex formulation using a nuclear-norm minimization technique [11]. These approaches are based on the assumption that the magnitude spectrogram of each source can be represented using a low-rank model.

To make it valid for the real-world situation, however, the following characteristics of the user annotation must be considered. First, the annotation is incomplete – that is, not all the T-F coefficients can be annotated. Moreover, the user cannot provide the exact target values, but only a binary annotation to indicate that the source is present or not for specific T-F regions. Finally, it is likely that some errors are present in the user annotation.

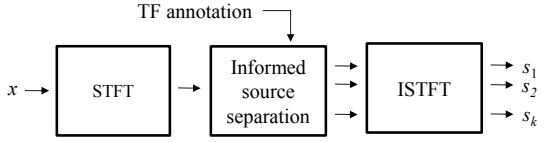


Fig. 1. Overall framework of the informed source separation with T-F annotation.

In this paper, we propose a novel ISS method to solve the abovementioned problems when a user-provided T-F annotation is limited. To this end, we first roughly set the target of each source from the binary annotation. With the low-rank model of each source using the nuclear-norm minimization, we also minimize the l_1-norm between the T-F coefficients of the source and the target to increase the robustness to the annotation error.

2. ISS WITH T-F ANNOTATION

2.1 Problem setting

In this section, we specify the problem that we aim to solve. Let us say we have an input mixture x, which is the sum of single sources s_k where k denotes the index of each source. The short-time Fourier transforms (STFTs) are then denoted as $\boldsymbol{X} \in \mathbb{C}^{F\times T}$ and $\boldsymbol{S}_k \in \mathbb{C}^{F\times T}$, respectively, where F and T denote the number of the frequency bins and the time frames, respectively. The magnitude spectrograms are denoted as $X = |\boldsymbol{X}|$ and $S_k = |\boldsymbol{S}_k|$, respectively. In addition, we define the target values of the k-th source as $T_{k,\Omega_k} = \{T_{k,f,t} \mid (f,t) \in \Omega_k\}$, where Ω_k is the set of the annotated T-F coefficients for the k-th source.

Unfortunately, it is practically impossible to expect the user to provide the exact target values T_{k,Ω_k}. We assume instead that what the user provides is just a binary annotation B_{k,Ω_k} where it is set to 1 if the user guesses the k-th source is 'present' in Ω_k, and 0 when it is 'absent'. We can thus roughly set T as follows based on B:

$$T_{k,\Omega_k} = \begin{cases} \dfrac{X_{\Omega_k}}{n_\Omega} & \text{if } B_{k,\Omega_k} = 1 \\ 0 & \text{if } B_{k,\Omega_k} = 0, \end{cases} \quad (1)$$

Algorithm 1. ISS with T-F annotation

Set $\rho = 1.1$ and $\mu = 0.01$.
Initialize E, Z, Λ, M, N = 0.
Until converge
  for k
    update S_k by (6) and (7).
    update $Z_{k,\Omega k}$ by (9).
    update E_k by solving (10).
  end
  for k
    update Λ_k by $\Lambda_k \leftarrow \Lambda_k + \mu(E_k - S_k)$.
    update M_k by $\mathrm{M}_k \leftarrow \mathrm{M}_k + \mu(Z_k - (S_k - T_k))$.
  end
  update N by $\mathrm{N} \leftarrow \mathrm{N} + \mu(X - \sum_k S_k)$.
  update μ by $\mu \leftarrow \rho\mu$.
end

where n_Ω is the number of k for which is $B_{k,\Omega_k} = 1$.

2.2 Proposed method

To estimate each s_k from x and T_{k,Ω_k}, we assume that the objective function must satisfy the following three necessary conditions:

a) *The difference between the obtained mixture and the sum of the estimated sources must be minimized.*
b) *The difference between the target and the estimated source values must be minimized.*
c) *The T-F representation of each source must be represented by a low-rank model.*

To make the problem simple, we handle all the sources and the mixture in the magnitude spectrogram domain, and thus a) and b) can be approximated as $\sum_k S_k \approx X$ and $S_{k,\Omega_k} \approx T_{k,\Omega_k}$, respectively. For c), we apply the concept of nuclear-norm minimization to approximately represent the low-rankness of the matrix. Based on the abovementioned conditions, we derive the objective function as follows:

$$\begin{aligned} \min_S \quad & \sum_k \|S_k\|_* + \lambda \sum_k \left\| S_{k,\Omega_k} - T_{k,\Omega_k} \right\|_1 \\ \text{s.t.} \quad & \sum_k S_k = X, \end{aligned} \tag{2}$$

where $\|\ \|_*$ and $\|\ \|_1$ denote the nuclear-norm and the l_1-norm, which are the sum of the singular values and the absolute values, respectively. λ is a parameter that controls the relative weights between the two terms. In this objective function, the conditions a) and b) are applied by means of a constraint and l_1-norm minimization. The effectiveness of this approach will be discussed in Section 2.4.

2.3 Augmented Lagrangian multiplier (ALM) method

The ALM method is often used to deal with the nuclear-norm and l_1-norm minimization such as in robust principal component analysis [12]. First, (2) can be equivalently rewritten as

$$\begin{aligned} \min_S \quad & \sum_k \|E_k\|_* + \lambda \sum_k \left\| Z_{k,\Omega_k} \right\|_1 \\ \text{s.t.} \quad & \sum_k S_k = X, E_k = S_k, Z_{k,\Omega_k} = S_{k,\Omega_k} - T_{k,\Omega_k}. \end{aligned} \tag{3}$$

According to the ALM method, we need to solve the following problem:

$$\begin{aligned} \min_{S,E,Z_\Omega} \quad & \sum_k \left(\|E_k\|_* + \lambda \left\| Z_{k,\Omega_k} \right\|_1 \right) \\ & + \frac{\mu}{2} \sum_k \left(\left\| E_k - S_k + \frac{1}{\mu} \Lambda_k \right\|_F^2 \right) \\ & + \frac{\mu}{2} \sum_k \left(\left\| Z_{k,\Omega_k} - \left(S_{k,\Omega_k} - T_{k,\Omega_k} \right) + \frac{1}{\mu} \mathrm{M}_{k,\Omega_k} \right\|_F^2 \right) \\ & + \frac{\mu}{2} \left\| X - \sum_k S_k + \frac{1}{\mu} \mathrm{N} \right\|_F^2, \end{aligned} \tag{4}$$

where $\Lambda_k \in \mathbb{R}^{F \times T}$, $\mathrm{M}_k \in \mathbb{R}^{F \times T}$, and $\mathrm{N} \in \mathbb{R}^{F \times T}$ are the ALMs. Since the number of variables is three or S, E, and $Z_{\Omega k}$, they are optimized alternately. Here we explain the optimization method for each variable, and the overall framework of the ALM method for (4) is described in Algorithm 1.

Solution for S: When E and $Z_{\Omega k}$ are fixed, (4) can be simplified as follows:

$$\begin{aligned} \min_S \quad & \sum_k \left(\left\| E_k - S_k + \frac{1}{\mu} \Lambda_k \right\|_F^2 \right) \\ & + \sum_k \left(\left\| Z_{k,\Omega_k} - \left(S_{k,\Omega_k} - T_{k,\Omega_k} \right) + \frac{1}{\mu} \mathrm{M}_{k,\Omega_k} \right\|_F^2 \right) \\ & + \left\| X - \sum_k S_k + \frac{1}{\mu} \mathrm{N} \right\|_F^2, \end{aligned} \tag{5}$$

and the optimal solution for S is given by

$$\begin{aligned} S_{k,\Omega} = \frac{1}{3} \Bigg(& \left(E_{k,\Omega_k} + \frac{1}{\mu} \Lambda_{k,\Omega_k} \right) + \left(Z_{k,\Omega_k} + T_{k,\Omega_k} + \frac{1}{\mu} \mathrm{M}_{k,\Omega_k} \right) \\ & + \left(X_{\Omega_k} + \frac{1}{\mu} \mathrm{N}_{\Omega_k} \right) - \sum_{\bar{k} \neq k} S_{\bar{k},\Omega_k} \Bigg), \end{aligned} \tag{6}$$

$$S_{k,\bar{\Omega}_k} = \frac{1}{2} \left(\left(E_{k,\bar{\Omega}_k} + \frac{1}{\mu} \Lambda_{k,\bar{\Omega}_k} \right) + \left(X_{\bar{\Omega}_k} + \frac{1}{\mu} \mathrm{N}_{\bar{\Omega}_k} \right) - \sum_{\bar{k} \neq k} S_{\bar{k},\bar{\Omega}_k} \right), \tag{7}$$

where $\bar{\Omega}_k = \{(f,t) \mid (f,t) \notin \Omega_k\}$.

Solution for $Z_{\Omega k}$: With S and E fixed, (4) is simplified as a function of $Z_{\Omega k}$, and can be expressed as

$$\min_{Z} \quad \sum_k \left\| Z_{k,\Omega_k} \right\|_1 + \frac{\mu}{2\lambda} \sum_k \left(\left\| Z_{k,\Omega_k} - \left(S_{k,\Omega_k} - T_{k,\Omega_k} \right) + \frac{1}{\mu} \mathrm{M}_{k,\Omega_k} \right\|_F^2 \right), \tag{8}$$

and the optimal solution for $Z_{\Omega k}$ is given by

$$Z_{k,\Omega_k} = \sigma_{\lambda/\mu} \left(S_{k,\Omega_k} - T_{k,\Omega_k} - \frac{1}{\mu} \mathrm{M}_{k,\Omega_k} \right), \tag{9}$$

where σ is a shrinkage operator $\sigma_\alpha(a) = \mathrm{sgn}(a) \max(|a| - \alpha, 0)$ and applied element-wisely for matrices.

Solution for E: With S and $Z_{\Omega k}$ fixed, (4) is simplified as a function of E, and can be written as

$$\min_{E} \quad \sum_k \left\| E_k \right\|_* + \frac{\mu}{2} \sum_k \left(\left\| E_k - S_k + \frac{1}{\mu} \Lambda_k \right\|_F^2 \right). \tag{10}$$

Candès and Li suggested that (10) can be solved by solving the following function [12]:

$$\min_{\Delta} \quad \sum_k \left\| \Delta_k \right\|_1 + \frac{\mu}{2} \sum_k \left(\left\| \Delta_k - \Phi_k \right\|_F^2 \right), \tag{11}$$

where Δ_k and Φ_k are the ordered singular value matrices of E_k and $S_k - \frac{1}{\mu}\Lambda_k$. The optimal Δ is given by

$$\Delta_k = \sigma_{1/\mu}(\Phi_k). \tag{12}$$

The optimal E_k can then be obtained as $Q_k \Delta_k R_k$, where $Q_k \Phi_k R_k$ is the singular value decomposition of $S_k - \frac{1}{\mu}\Lambda_k$.

2.4 Comparison with Lefèvre's method

Recently, Lefèvre *et al*. proposed the similar method [11], and it is worth to compare it with the proposed method. Basically, Lefèvre's method aims to solve the following objective function[1].

$$\begin{aligned} \min_{S} \quad & \sum_k \left\| S_k \right\|_* + \eta \left\| X - \sum_k S_k \right\|_F \\ \text{s.t.} \quad & S_{k,\Omega_k} = T_{k,\Omega_k} \text{ for all } k, \end{aligned} \tag{13}$$

where η is a parameter for the relative weights. It is noted that the objective functions (2) and (13) are quite similar, and both contain all the conditions described in 2.2. However, the condition b) is strictly constrained in (13), or $S_{k,\Omega_k} = T_{k,\Omega_k}$, whereas in (2) it is given as one of the terms in the objective function, *i.e.*, $\sum_k \left\| S_{k,\Omega_k} - T_{k,\Omega_k} \right\|_1$, which allows tolerance.

The advantage of the proposed method is clear when the user makes some errors in the annotation, because (13) does not allow any tolerance. In case of (2), the errors or the false annotations are considered as outliers, making the low-rank modeling difficult, and thus are ignored by l_1-norm minimization. The effectiveness of the proposed method will be discussed more in the next section.

3. EXPERIMENTS

3.1 Dataset

In order to evaluate our algorithm, we used the SISEC database [13], which contains five pieces of professional music recordings of 10-25s length. Because this database was recorded in a multi-track format, we mixed each multi-track recording down to a two-track format of the vocal (k=1) and the accompaniment (k=2), where the latter contains all the non-vocal sources. All the tracks were resampled to 16kHz to reduce the computational burden. Finally, we mixed them so that the input signal should have 0dB and -10dB vocal-to-accompaniment ratio (VAR) to compare the robustness of the algorithms.

3.2 Evaluation

We calculated the signal-to-distortion ratio (SDR) using the BSS-EVAL 3.0 as an evaluation metric [14]. To evaluate the robustness under various mixing and annotation conditions, we experimented with the various annotation rates (ARs), and the false annotation rates (FARs), which are defined as

$$\mathrm{AR} = \frac{\sum_k \text{the number of the annotated coefficients of } S_k}{\sum_k \text{the total number of the coefficients of } S_k},$$

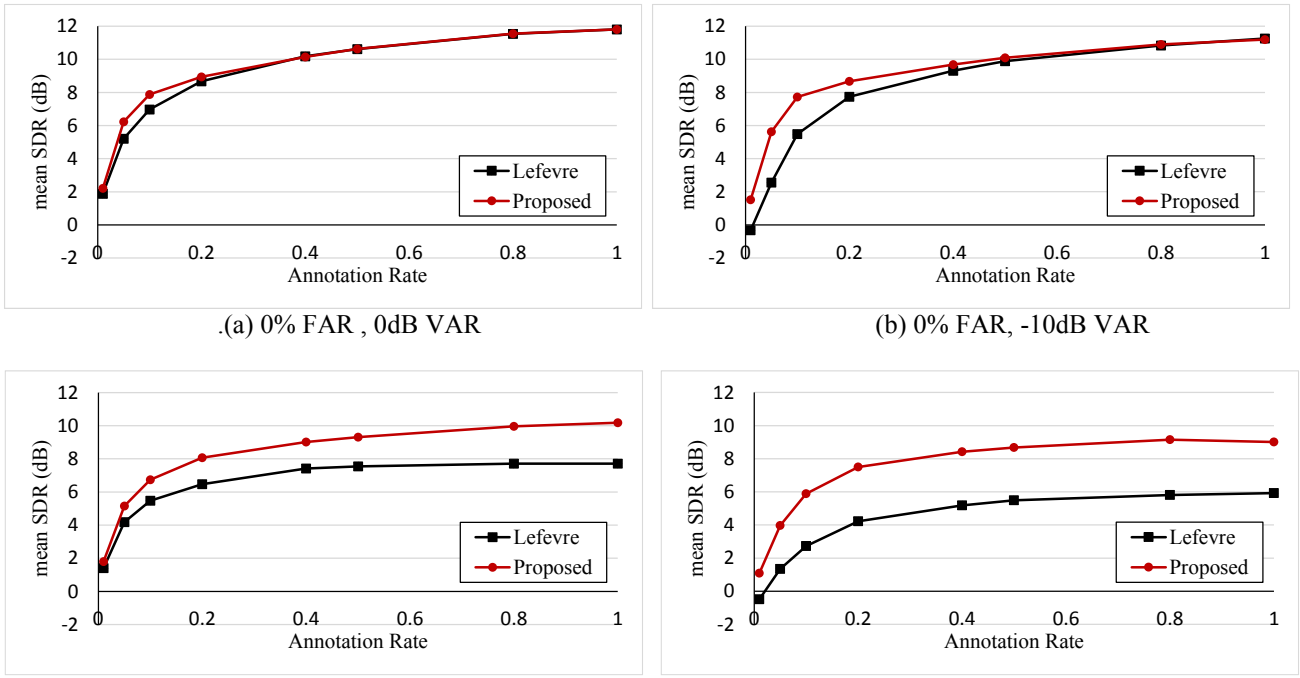
$$\mathrm{FAR} = \frac{\sum_k \text{the number of the false-annotated coefficients of } S_k}{\sum_k \text{the number of the annotated coefficients of } S_k},$$

where 'false annotation' denotes $B_{k,\Omega_k} = 1$ when the source is actually absent, or $B_{k,\Omega_k} = 0$ when it is actually present. In general, the separation task would become easier under the high AR and the low FAR conditions.

3.3 Experiment settings

For the T-F representation, we used the STFT with the hamming window of 512 samples and the 256-sample overlap. The annotation is provided by randomly selecting the T-F coefficients with the probability of AR. $B_{k,\Omega}$ is set to 1 if $S^*_{k,\Omega_k} \geq S^*_{\bar{k} \neq k, \Omega_{\bar{k}}}$, and 0 otherwise, where S^* is the magnitude spectrograms of actual source. For the normalization factor of λ, we adopted $\lambda = \lambda' / \sqrt{\max(F,T)}$, as suggested by Candès and Li [12].

[1] To be precise, Lefèvre's method solves the problem using the power spectrograms, while we use the magnitude spectrograms. In our experiments in Section 3, however, we used the magnitude spectrograms for Lefèvre's method as well because they achieve the better performance.



.(a) 0% FAR , 0dB VAR

(b) 0% FAR, -10dB VAR

(c) 10% FAR, 0dB VAR

(d) 10% FAR, -10dB VAR

Fig. 2. Comparison between the Lefèvre's [11] and the proposed methods.

Table 1. Performance as a function of λ' for different FAR conditions. AR is set to 20%.

mean SDR		λ'						
		0.1	**0.2**	**0.5**	**1**	**2**	**5**	**10**
FAR (%)	**0**	3.1	4.7	7.2	8.3	8.9	9.3	9.3
	10	2.5	3.8	6.0	7.5	8.0	7.4	6.9
	20	1.9	2.8	4.6	5.9	6.2	5.2	4.7

Table 2. Performance as a function of λ' for different AR conditions. FAR is set to 10%.

mean SDR		λ'						
		0.1	**0.2**	**0.5**	**1**	**2**	**5**	**10**
AR (%)	**10**	1.6	2.5	4.3	5.8	6.7	6.6	6.0
	20	2.5	3.8	6.0	7.5	8.0	7.4	6.9
	40	3.7	5.5	7.8	8.7	8.9	8.4	7.8

3.4 Results

First, we examined the effect of the weight parameter λ. As mentioned in Section 3.3, we decided the normalization factor and tried to find the optimal λ'. Table 1 and 2 show the separation performance as a function of λ'. The results from Table 1 indicate that with too large λ' the performance degrades when the annotation is not correct. This makes sense because λ' determines how close the separated source is to the target, and therefore it causes a negative impact when the target is not exact. On the other hand, Table 2 shows that λ' does not affect the performance much with regard to AR. From these experiments, we set the value of λ' to 2.

For the next experiment, we compared our method to the Lefèvre's. We set the value of η to 0.1 as the authors proposed [11]. Fig. 2 illustrates the mean SDR of the two methods. When a user-guided annotation is complete and exact (*i.e.*, 100% AR, 0% FAR), both algorithms achieve high SDR (Fig.2(a)), but the proposed algorithm is slightly better when the energy difference between the sources gets larger (Fig.2(b)). With errors present in the annotation or nonzero FAR, however, the proposed method yields significantly higher SDR (Fig.2(c)), and the performance gap gets even wider when the VAR becomes lower (Fig.2(d)). In particular, comparing the easiest case (Fig.2(a)) with the hardest one (Fig.2(d)), the performance decrease is much less severe with the proposed method (8.94dB → 7.50dB at 20% AR) than with the state-of-the-art (8.67dB → 4.22dB). These results imply that the proposed algorithm is significantly more robust under real-world situations.

4. CONCLUSION

In this paper, we proposed a novel method for informed source separation using a low-rank model, assuming incomplete T-F annotations are given by a user. In particular, we showed that the use of l_1-norm, which allows the errors from the annotation, can dramatically increase the robustness of the separation algorithm. For future work, we plan to generalize the optimization framework by using the Schatten p-norm and the l_p-norm. Moreover, we will extend the proposed method to handle multi-channel signals.

5. REFERENCES

[1] P. Smaragdis and G.J. Mysore, "Separation by "humming": User-guided sound extraction from monophonic mixtures," In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, NY, 2009.

[2] D. FitzGerald, "User assisted source separation using non-negative matrix factorization," In *Proc. IET Irish Signals and Systems Conference*, Dublin, Ireland, 2011.

[3] P. Hanna and M. Robine, "Query by tapping system based on alignment algorithm," In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, 2009.

[4] S. Ewert and M. Muller, "Using score-informed constraints for NMF-based source separation," In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, 2012.

[5] J.-L. Durrieu and J.-P. Thiran, "Musical audio source separation based on user-selected f0 track," In *Proc. International Conference on Latent Variable Analysis and Signal Separation*, Tel-Aviv, Israel, 2012.

[6] L.L. Magoarou, A.Ozerov, and N.Q.K. Duong, "Text-informed audio source separation using nonnegative matrix partial co-factorization," In *Proc. IEEE International Workshop on Machine Learning for Signal Processing*, Southampton, UK, 2013.

[7] N.J. Bryan and G.J. Mysore, "Interactive refinement of supervised and semi-supervised sound source separation estimates," In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, 2013.

[8] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. International Conference on Independent Component Analysis and Signal Separation*, London, UK, 2007.

[9] A. Lefèvre, F. Bach, and C. Févotte, "Semi-supervised NMF with time-frequency annotations for single channel source separation," In *Proc. International Society for Music Information Retrieval Conference*, Proto, Portugal, 2012.

[10] D.D. Lee and H.S. Seung, "Algorithms for Non-negative Matrix Factorization," *Neural Information Processing Systems*, MIT Press, pp. 556–562, 2001.

[11] A. Lefèvre, F. Glineur, and P.-A. Absil, "A convex formulation for informed source separation in the single channel setting," *Neurocomputing*, pp. 26-36, 2014.

[12] E.J. Candès and X. Li, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, May 2011.

[13] N. Ono, Z. Koldovský, S. Miyabe, and N. Ito, "The 2013 signal separation evaluation campaign," In *Proc. IEEE International Workshop on Machine Learning for Signal Processing*, Southampton, UK, 2013.

[14] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.